Early pharmacokinetic optimisation is a key principle in drug discovery and development. Modeling absorption, distribution, metabolism and excretion (ADME) using experimentally-derived data is time-consuming and expensive. The use of computational in silico techniques to predict pharmacokinetic properties based on molecular structure is gaining wider validity and acceptance in the pharmaceutical industry. This book describes the use of artificial neural networks (ANN) as robust nonlinear modeling tools for developing quantitative structure-pharmacokinetic relationships (QSPkR). Different ANN paradigms are examined for predictive modeling of various pharmacokinetic parameters, both individually and simultaneously. Consideration is given to physiological processes, drug and molecular structural data, and model interpretation. As well as providing the theory behind ANN model construction, this book details their practical application in pharmaceutical research and gives meaning to many of the theoretically-derived molecular descriptors now available.

A valuable resource for medicinal chemists and pharmaceutical scientists engaging in structure-property and structure-activity modeling.

**Joseph Turner, Snezana Agatonovic-Kustrin**

MBBS, University of Queensland, Australia; BMedSc (Hons), University of Sydney, Australia; PhD (Pharmacy), University of Sydney, Australia; Conjoint Lecturer, School of Medicine, University of Queensland, Australia; Rural Generalist, Queensland Health, Australia.

Snezana Agatonovic-Kustrin: BSc (Pharmacy), University of Belgrade, Serbia; MSc (Pharmacy), University of Belgrade, Serbia; PhD (Pharmaceutical Chemistry), University of Belgrade, Serbia; Senior Lecturer (Pharmaceutical Sciences), School of Pharmacy and Molecular Sciences, James Cook University, Australia.

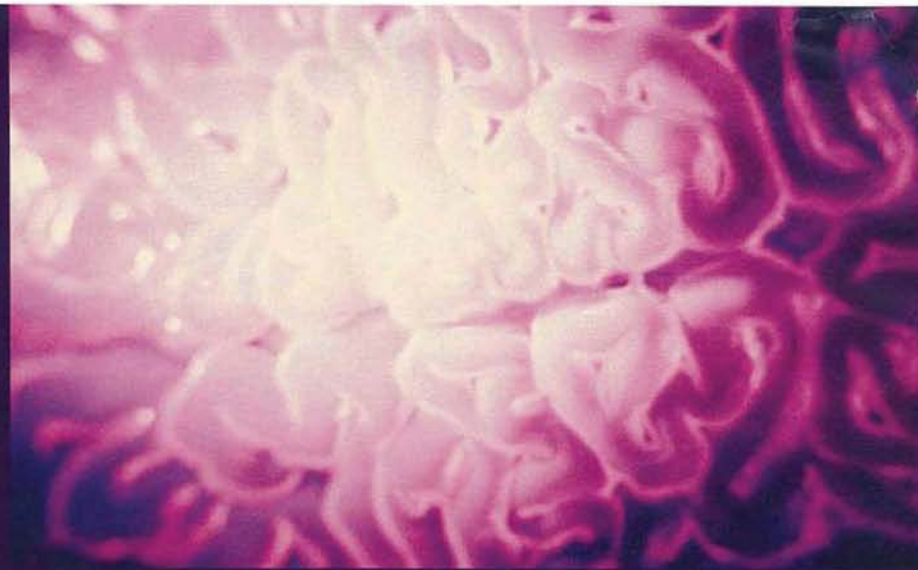
9 783836 480383

978-3-8364-8038-3

Joseph Turner, Snezana Agatonovic-Kustrin

Structure-Pharmacokinetic ANN Modeling

VDM



**Joseph Turner,**
**Snezana Agatonovic-Kustrin**

# Quantitative Structure-Pharmacokinetic Relationships

## Artificial Neural Network Modeling

VDM Verlag Dr. Müller

Joseph Turner,
Snezana Agatonovic-Kustrin

# Quantitative Structure-Pharmacokinetic Relationships

Joseph Turner,
Snezana Agatonovic-Kustrin

# Quantitative Structure-Pharmacokinetic Relationships

## Artificial Neural Network Modeling

VDM Verlag Dr. Müller

## Imprint

Bibliographic information by the German National Library: The German National Library lists this publication at the German National Bibliography; detailed bibliographic information is available on the Internet at http://dnb.d-nb.de.



Cover image: www.purestockx.com

Publisher:
VDM Verlag Dr. Müller Aktiengesellschaft & Co. KG, Dudweiler Landstr. 125 a, 66123 Saarbrücken, Germany,
Phone +49 681 9100-698, Fax +49 681 9100-988,
Email: info@vdm-verlag.de



Produced in USA and UK by:
Lightning Source Inc., La Vergne, Tennessee, USA
Lightning Source UK Ltd., Milton Keynes, UK
BookSurge LLC, 5341 Dorchester Road, Suite 16, North Charleston, SC 29418, USA

**ISBN: 978-3-8364-8038-3**

# Table of Contents

# Acknowledgements

Part of this work was drawn from the dissertation submitted by Joseph V Turner for his PhD (2004) at the Faculty of Pharmacy, The University of Sydney, Australia. His supervisors over that time are gratefully acknowledged: Dr Desmond J Maddalena, Dr Snezana Agatonovic-Kustrin, Dr David J Cutler, and Prof Hak-Kim Chan.

The following acknowledgments are made where material has been subject to peer-review and published elsewhere:

Chapter 3. Reproduced from the *Journal of Compationtional Chemistry*, Turner JV, Cutler DJ, Spence I and Maddalena DJ. Selective descriptor pruning for QSAR/QSPR studies using artificial neural networks, 24(7):891-897, copyright (2003), with permission from John Wiley & Sons.

Chapter 4. Reproduced from the *International Journal of Pharmaceutics*, Turner JV, Maddalena DJ, Cutler DJ, Pharmacokinetic parameter prediction from drug structure using artificial neural networks, 270(1-2): 209-219, copyright (2004) with permission from Elsevier.

Chapter 5. Reproduced from the *Journal of Pharmaceutical Sciences*. Turner JV, Maddalena DJ, Cutler DJ and Agatonovic-Kustrin S. Multiple pharmacokinetic parameter prediction for a series of cephalosporins. 92(3): 518-525, copyright (2003) with permission from John Wiley & Sons.

Chapter 6. Reproduced from *Pharmaceutical Research*, Turner JV, Maddalena DJ and Agatonovic-Kustrin S, Bioavailability prediction based on molecular structure for a diverse series of drugs, 21(1): 68-82, copyright (2004), with permission from Springer.

Reproduced from *Analytica Chimica Acta*, Turner JV, Glass BD and Agatonovic-Kustrin S, Prediction of drug bioavailability based on molecular structure, 485(1): 89-102, copyright (2003) with permission from Elsevier.

Chapter 7. Reproduced from *Drug Metabolism Letters*, Agatonovic-Kustrin S, Turner JV, and Glass BD, Quantitative structure-retention-pharmacokinetic relationship studies, 2(2): 130-137, copyright (2008) with permission from Bentham.

# List of Figures

# List of Tables

# Abbreviations

| | |
|---|---|
| 1D | one-dimensional |
| 2D | two-dimensional |
| 3D | three-dimensional |
| ADME(T) | absorption, distribution, metabolism, excretion, (toxicology) |
| AI | artificial intelligence |
| ANN | artificial neural network |
| AS | artificial structured |
| $C_A$ | arterial drug concentration |
| $C_b$ | concentration of drug in body |
| $CL$ | clearance (total) |
| $CL_{NR}$ | clearance (nonrenal) |
| clog $P$ | calculated log $P$ |
| $CL_R$ | clearance (renal) |
| $C_p$ | concentration of drug in plasma |
| $C_t$ | concentration of drug in tissue |
| $C_V$ | venous drug concentration |
| CYP | cytochrome P450 |
| $D_b$ | amount of drug in the body |
| ER | efficiency ratio |
| $ER$ | extraction ratio |
| $f_b$ | fraction bound to plasma proteins |
| $f_e$ | fraction excreted in urine |
| GA | genetic algorithm |
| GI | gastrointestinal |
| GNN | genetic neural network |
| GRNN | generalised regression neural network |
| HIA | human intestinal absorption |
| HLB | hydrophilic-lipophilic balance |
| HOMO | highest occupied molecular orbital |
| HT | high-throughput |
| ICRMW | inverse cube-root of molecular weight |
| ISRMW | inverse square-root of molecular weight |
| $k'$ | retention |
| LFER | linear free energy related |
| LOO | leave-one-out |
| LUMO | lowest unoccupied molecular orbital |
| MLP | multilayer perceptron |
| MLR | multilinear regression |
| MR | molar refractivity |
| MW | molecular weight |
| NCE | new chemical entity |
| PBPK | physiologically-based pharmacokinetic |
| PCA | principal component analysis |
| PSA | polar surface area |
| $P_{t:b}$ | tissue/blood partition coefficient |
| $P_{t:p}$ | tissue/plasma partition coefficient |
| $Q$ | blood flow |

| | |
|---|---|
| QSAR | quantitative structure-activity relationship |
| QSPkR | quantitative structure-pharmacokinetic relationship |
| QSPR | quantitative structure-property relationship |
| $r_{cv}$ | cross-validation correlation |
| RBF | radial-basis function |
| RMS | root mean squared |
| $r_t$ | training correlation |
| $r_{tes}$ | testing correlation |
| $r_{val}$ | validation correlation |
| S:N | signal to noise ratio |
| SD | standard deviation |
| SOM | self-organising map |
| SAR | structure-activity relationship |
| SPR | structure-property relationship |
| SWR | stepwise regression |
| $t_{1/2}$ | half life |
| $V_{SS}$ | volume of distribution at steady state |
| WDI | World Drug Index |
| $w_{meaningful}$ | meaningful weight value |
| $w_{meaningless}$ | meaningless weight value |

# Chapter 1. Introduction

## 1.1 General Matter

The technology and research revolution has provided many areas of science and industry with tools for more extensive and efficient operation. Nowhere is this phenomenon more evident than for new drug discovery and development in the pharmaceutical industry. Exploring the relationship between the structure of a molecule and its various biological and biochemical properties is the basis of drug discovery. Modern approaches to this field of study employ a combination of techniques. These include tests based on combinatorial chemistry and high-throughput (HT) screening as well as rational pharmaceutical design based on geometric and chemical characteristics of molecule-molecule interactions. Furthermore, understanding and optimising factors such as the effect of a compound on the body and the effect of the body on a compound are essential in developing a new drug.

The main bottleneck in drug discovery is the identification of new chemical entities (NCEs) to be used for drug leads. The 1990s saw development of new automated tools for drug discovery including combinatorial chemistry and high-throughput screening. These tools have led to the increased discovery of new drug lead compounds each of which in turn require pharmacological and pharmacokinetic testing. Moreover, substantial increases in computing power as well as development of robust software has given scientists the opportunity to undertake significant research projects from their own desktops. Consequently, data analysis, data mining, and information manipulation have all benefited and progressed considerably.

Software programs have been developed for a wide range of fields such as quantitative structure-activity relationship (QSAR) studies, pharmacophore elucidation, molecular modeling, drug-receptor interactions and *in vivo* simulations. Newer techniques have been influenced by what is termed "soft computing" which aims to accommodate the imprecision and uncertainty inherent in the real world [Zadeh, 1996]. Soft computing draws on the model of the human brain and derives mainly from artificial intelligence (AI) sources including genetic algorithm (GA), fuzzy logic, and artificial neural network (ANN) approaches [Maddalena, 1998]. Other less common techniques include cellular automata, fractals and chaos theory. ANNs are aparticularly useful modeling tools for nonlinear systems. Although not as common in the pharmaceutical industry as conventional modeling and mathematical techniques, soft computing has been successful in a number of fields in the industry.

In particular, soft computing has been useful in the development of quantitative structure-activity relationship and quantitative structure-property relationship (QSPR) models. General methods involve correlation of physicochemical descriptors of chemical compounds with either an activity or property value. Classically, one or two descriptors such as octanol/water partition coefficients and molar refractivity are experimentally determined for a group of congeneric compounds [Hansch et al., 1995]. These experimental descriptor values are then related to selected biological activity. The result is a mathematical model which describes the contribution of the descriptors to the activity. Once a predictive model has been built, numerous new potential-drug molecules which are chemically similar to those of the benchmark dataset can then be screened from large databases. These molecules are all evaluated for their biological properties based on the predictive model developed. The aim is to target a few novel molecules with potentially attractive pharmaceutical properties that can then be tested further in the traditional way in the laboratory. Effective data mining techniques are vital to extract the information necessary to select these novel molecules.

Such models have used both whole molecule descriptors and descriptors for individual substitution positions and functional groups on structurally related compounds. Thus, for a single study there may be a large number of potential descriptors for correlation with only a single target activity or property. Not all descriptors are useful so selection of meaningful descriptors is crucial for successful model development.

Recently, theoretical descriptors generated only from the molecular structure of a compound have become popular. Over a thousand of these descriptors have been defined to date although not all are entirely useful [Balaban & Ivanciuc, 1999]. Many of these descriptors have been successfully correlated with parameters such as boiling points of alkanes [Cherqaoui & Villemin, 1994], aqueous solubility [Huuskonen et al., 1997], binding affinities [Beck et al., 1996], and analgesic properties [Galvez et al., 1994b].

Similarly, quantitative structure-pharmacokinetic relationship (QSPkR) models have also been constructed to correlate drug stuctures and their pharmacokinetic parameters [Seydel & Schaper, 1981; Hinderling et al., 1984a; Fouchecourt et al., 2001; Agatonovic-Kustrin et al., 2008]. However, QSPkRs are not as common in the literature as other QSPRs are. This has largely been attributed to the complex factors involved in some pharmacokinetic parameters such as hepatic metabolism and elimination half life, as well as the time dependency of drug concentration *in vivo* [Mayer & van de Waterbeemd, 1985].

The development of predictive QSPkR models are of increasing interest to the pharmaceutical industry since it would allow valuable information to be gained very early during the drug discover/development process. Should successful models be based on structure alone then predictions could be made for theoretical chemical structures during screening before they are even synthesised. Additionally, knowledge of human pharmacokinetics prior to clinical trials would enable decisions to be made regarding viability of potential drugs for continued development. This would impart another substantial time- and cost-saving benefit.

There are three essential components involved in structure-pharmacokinetic modeling: 1) acquisition of pharmacokinetic data; 2) generation of theoretical molecular descriptors; and 3) model construction. It is difficult to gather consistent pharmacokinetic data from the literature since utilisation of different sources inevitably leads to increased variability amongst the data. There is also a huge variety of molecular descriptors and, not surprisingly, different authors rarely use the same descriptors. Given that this modeling technique is data-driven, the nature and meaning of descriptors selected in models is also very important.

The remainder of this introduction will deal with the relevant aspects of drug discovery and development, pharmacokinetics, pharmakokinetic modeling techniques, available descriptors, and a review of current QSPkR modeling.

### 1.1.1 Objectives

The broad aim of this work was to develop predictive QSPkR models using ANNs. Specific objectives were to:

- Identify various theoretical descriptors generated from molecular structure and examine their relevance to QSPkR studies.
- Investigate both selection of descriptors and effects of ANN architecture on QSPkR model performance.
- Explore relationships of theoretical descriptors with different pharmacokinetic parameters.
- Investigate the viability of multiple, simultaneous pharmacokinetic parameter prediction.

- Develop predictive QSPkR models for both structurally related and structurally unrelated sets of drugs.
- Examine models with both theoretical and experimentally-derived descriptors.

## 1.2 Pharmaceutical Product Development

Development of successful pharmaceutical products drives profit, which in turn permits further drug development. Profit from sales occurs only after a marketable product has been produced and its associated developmental costs have been surpassed. In addition, sales for a successful product must also account for the cost of unsuccessful compounds that have failed at some stage during development.

### 1.2.1 An Overview

The drug discovery and development process involves elements from both regulatory bodies and industry (Figure 1-1). The entire process spans pre-clinical laboratory research and development, through to clinical evaluation, and finally to post-marketing surveillance. It is in pre-clinical research, commonly referred to as Phase 0, that new drug entities are screened and developed for eventual clinical application to humans. Owing to rapid advances in areas such as computing technology, combinatorial chemistry, molecular and cell biology, and high-throughput screening techniques, strong progress has been made in identifying potential lead compounds. Although such techniques have provided an increased number of potential new drug entities, this has not necessarily translated to an increased number of drugs successfully reaching the marketplace [Grass & Sinko, 2001].

Of all the NCEs screened in Phase 0, only a small number ever progress beyond animal studies. From this small percentage it has been estimated that less than one quarter possess all the necessary pharmacokinetic and pharmacodynamic characteristics to successfully become marketable products. Increasing the number of NCEs progressing to the clinical trial phases then substantially increases the number of failures at this late stage. The bulk of drug development spending can be attributed to these failures, and the total amount has been estimated to be around 75% of all monies spent on drug development. Hence, the focus of drug development has expanded more and more to include procedures aimed at identifying potential failures as well as successes [Ekins et al., 2000].

In the period 1968-1988, it was found that the major reasons for failure of NCEs in humans were unacceptable pharmacokinetics (~40% of failures) and lack of efficacy (~30% of failures) [Prentis et al., 1988]. Unacceptable pharmacokinetics can include poor absorption, distribution, metabolism, or excretion (ADME) characteristics. Poor pharmacokinetics can also manifest as lack of clinical efficacy. Hence, the human pharmacokinetics of a compound plays a vital role in determining the suitability of an NCE for further development.

Screening for ADME properties and toxicity is usually performed both *in vitro* and with various animal models which are time-consuming and expensive [Norris et al., 2000]. Even then, results may not always accurately reflect the pharmacokinetics of a compound once it is administered to humans. Results gained from ADME screening are used to determine whether development of an NCE should continue to Phase 1 or not.

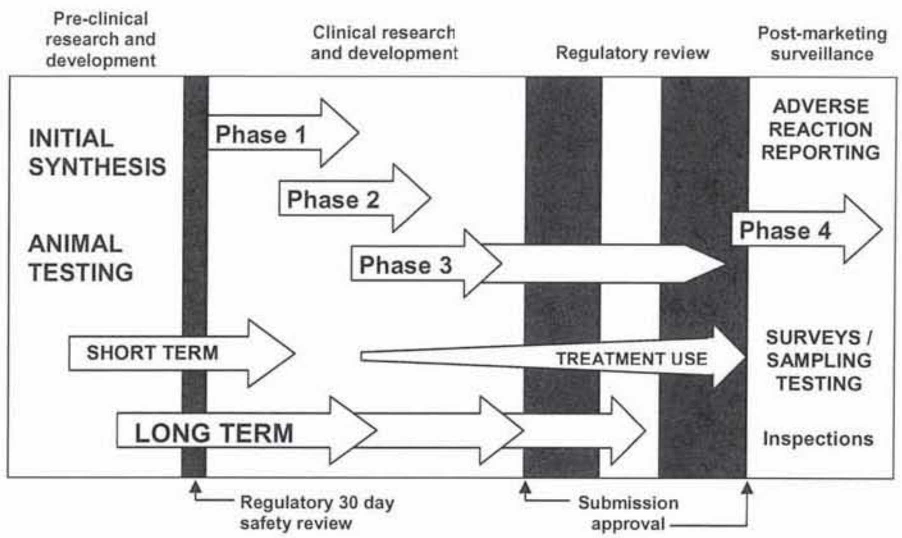



**Figure 1-1.** General sequence of events involved in drug development (adapted from [Lesko et al., 2000]). Black areas (■) represent time taken for regulatory processes, whilst clear areas represent scientific process normally associated with industry.

Early clinical testing in humans conducted in healthy subjects is aimed at determining tolerated dose size, initial pharmacokinetic profiles, candidate delivery systems, and the relationship between plasma concentrations and pharmacological effects [Peck et al., 1992]. Preclinical screening may indicate an NCE with suitable pharmacokinetic attributes, however, the majority of candidates do not succeed through the clinical testing phases.

Owing to financial pressures and the need for more accurate predictive methods, research into methods other than *in vitro* screening and animal models is a growing area. Computational techniques, often termed *in silico* methods, have begun to play a larger role in the drug discovery process. Previously impossible computational tasks have become a matter of routine with the ever-increasing power and availability of computers and software. *In silico* methods have gained popularity in virtual compound library screening [Walters et al., 1998], three-dimensional (3D) pharmacophore elucidation [Terfloth & Gasteiger, 2001], and QSPR analyses [Ekins et al., 2000]. They are cheaper and quicker than performing *in vitro* and animal experiments, although they are not yet as acceptable to regulatory authorities. Data taken from previously published work or from objective sources such as molecular structure have allowed prediction of important pharmacokinetic properties such as intestinal absorption [Wessel et al., 1998], distribution parameters [Herman & Veng-Pedersen, 1994], and binding characteristics [Wagener et al., 1995; Loukas, 2001]. Both experimental and *in silico* techniques have advantages and disadvantages, and their roles in drug development will now be discussed briefly.

### 1.2.2 *In Vitro* Screening and Animal Models

The use of *in vitro* and animal models allows research to be performed on a much cheaper and faster scale than in humans. Furthermore, safety issues and ethical requirements for humans are also avoided. These methods aim to provide measurements regarding the potential NCEs which can then hopefully be correlated with human activity or pharmacokinetics.

Most drugs are developed for oral administration so bioavailability of a compound is an important factor to consider. Oral bioavailability is dependent sequentially upon dissolution in the gastrointestinal (GI) tract, absorption across the physical barrier of the GI membrane, and a first pass through the liver and lungs [Sietsema, 1989]. Absorption occurs via either paracellular or transcellular pathways. Hence, absorption is influenced by permeability, molecular size, and hydrogen bonding characteristics [Smith & van de Waterbeemd, 1999]. Although not a measure of bioavailability, absorption is the first step in delivering an oral dose of a drug.

One common technique for *in vitro* modeling of absorption is the Caco-2 monolayer system which is an immortalised human colon adenocarcinoma cell line. Caco-2 cells are enterocyte-derived cells possessing a microvillus surface and allow moderate to high-throughput screening of compounds. Another technique, Madin-Darby canine kidney (MDCK) cells in monolayer can be used in a similar functional manner to Caco-2 cells but do not require the 2-3 week culturing times Caco-2 cells do [Pelkonen et al., 2001].

Although both techniques allow reasonably fast screening of compounds, the major disadvantage is that they are considerably different from the situation *in vivo*. Compounding this are the inter-experiment and inter-laboratory variations seen with these *in vitro* techniques. Thus, cellular models can provide useful information but are not complete in themselves.

High-throughput screening is the rapid analysis of chemical libraries for biological activity. Compounds are screened using automated miniaturised assays which enable vast numbers of compounds to be tested in a short period of time [Inglese, 2002].

Combinatorial libraries contain from hundreds to millions of compounds for testing, so examining the entire chemical space available to most companies would not be affordable. Efforts have instead been aimed at reducing the chemical space to those compounds for which manufacture and further development is a feasible proposition [Gobbi & Poppinger, 1998]. Searching such large numbers of compounds has increased the number of drug-like "hits" exhibiting potential biological activity. This has in turn placed more pressure on the subsequent step of assessing the potential NCEs for suitable pharmacokinetic properties.

*In vivo* metabolism of test compounds is a serious problem in new drug develoopment. Metabolism of compounds by various enzymes can also be screened using *in vitro* high-throughput methods. Typically, hepatic microsomal or other liver or tissue homogenate preparations are incubated in 96-well plates with individual compounds in each well. Reaction times and conditions are completely controlled, and extent of metabolism is compared with a standard.

The aim to mimic *in vivo* metabolism is not always fully achieved since drug metabolism is a multifactorial process which usually involves multiple pathways [Gaviraghi et al., 2001]. In addition, protein binding can limit hepatic extraction *in vivo* and may not be accounted for using simply the *in vitro* screen. The major drawback with microsomal preparations, however, is the inconsistency between preparations which can lead to variable results [Spalding et al., 2000].

Allometric scaling is another method of screening potential NCEs for human application. It is based on the premise that human and animal anatomical, physiological and biochemical characteristics are comparable [Feng et al., 2000]. Once pharmacokinetic parameters have been determined in animals they can then be mathematically related to human pharmacokinetic parameters. Individually, there is no single animal which can be relied upon to accurately predict human pharmacokinetics.

Consequently, data must often come from several different species to construct a one predictive model [Hussain et al., 1993]. Correction terms can also be applied to allometric calculations to increase the accuracy of models. Allometric scaling has been applied to structurally diverse compounds [Feng et al., 2000] and for various pharmacokinetic parameters [Jezequel, 1994].

Although allometric studies avoid experiments on humans, they are still subject to ethics approval and are expensive to run. As a result of the inherent differences between animals and humans, allometric models do not always lead to accurate predictions for a human clinical situation [Grass & Sinko, 2001].

### 1.2.3 *In Silico* Screening

Computational methods for ADME began with the classical QSAR models developed last century using lipophilicity [Hansch & Lien, 1968]. Even though datasets were small and comprised of structurally similar compounds, the hypothesis that metabolism and activity could be modeled in a quantitative fashion based on structural considerations was pioneering. From these beginnings, much progress has been made using purely computational methods for compound screening. The scope of QSAR has evolved to account for the spatial arrangements of atoms in a candidate molecule. Interaction of a molecule with a receptor or enzyme is dependant upon the 3D configuration and conformation of that molecule: good steric arrangement of atoms and functional groups allows a more positive interaction. Once a 3D pharmacophore has been generated for a known agonist or substrate, screening of potential candidates for that particular receptor or enzyme may then commence. Other computational techniques have been combined with 3D QSAR, for example comparative molecular field analysis (CoMFA) which estimates steric and electrostatic interactions between a molecule and target binding site, and the VolSurf/GRID procedure which calculates energy-favourable sites around a molecule and converts them into selected molecular descriptors [Ekins et al., 2000].

One of the critical requirements for these *in silico* screening techniques is the availability of virtual libraries of compounds containing structures able to be screened. If the number of compounds available for *in vitro* high-througput screening is vast, then the number of theoretical structures possible in virtual libraries is almost incomprehensible. A maximally diverse virtual library is neither practical nor possible, so attempts must be made to limit the size of the library. Compounds should be selected first on the basis of being synthesisable products: lengthy, expensive and low yield reactions would most likely not produce an economically viable product. Another method may be to restrict compounds to structures approaching those of current marketed drug products. The majority of commercial drugs can be represented by a limited number of structural or functional scaffolds [Bemis & Murcko, 1996].

A scaffold is the basic structural element used as the starting point for the generation of chemical libraries through chemical modification. There are two basic scaffold types, the functional scaffold and the structural scaffold. The functional scaffold is a molecule with specfic biological activity directed toward a molecular target. For example, only chemical structures known to be susceptible to a particular enzyme may be included. A functional scaffold is a lead compound used for the generation of focused libraries to optimise specific lead properties (eg potency, selectivity, or bioavailability) while maintaining the basic activity.

Because active molecules are the uilimale goal of every lead or drug discovery project, libraries based on functional scaffolds with proven activity are the most useful. The structural scaffold is a molecule with certain structural features (eg specific ring systems, chiral centres, or functional groups). Structural scaffold libraries may be more complementary to an existing generic library for hit and lead search over a wide range of target classes. These libraries can increase available chemical space and fill structural gaps and may therefore be more universally applicable than functional scaffold-based libraries.

Three types of libraries need also be defined: general, focussed, and targeted. In order of specificity, general libraries are the least specific and are designed to be arbitrarily of broad interest in high-througput screening. Focussed libraries are aimed at a family of related targets, for example cytochrome P450 (CYP) 2C9 substrates, whereas targeted libraries are specific for one particular binding or activity endpoint [Walters et al., 1998].

Such libraries may also be screened for desirable structural characteristics. It is known that hydrogen bonding, lipophilicity, and molecular surface properties can affect drug transport and membrane permeation [Stenberg et al., 2000]. These factors play an important role in bioavailability and, hence, suitability of a drug for manufacture as an oral formulation. Simple methods can be employed such as counting the number of hydrogen donors and acceptors, although more complex methods may be required for calculation of lipophilicity and molecular surface properties. The "Rule of Five" has been employed as a general guideline in industry to limit the size of virtual libraries to compounds likely to be adequately absorbed from the intestine [Lipinski et al., 1997]. The rule was developed upon examination of 2245 drugs from the World Drug Index (WDI) that were believed to have entered Phase II trials and were orally absorbed. According to the rule, compounds are deemed to have poor intestinal absorption if any two of the following conditions are met:

- ⇒ There are more than five hydrogen-bond donors.
- ⇒ The calculated log $P$ (clog $P$) is greater than five.
- ⇒ The molecular weight (MW) is over 500.
- ⇒ There are more than 10 hydrogen-bond acceptors.

Oxygen and nitrogen atoms are defined as being hydrogen bond acceptors, and –NH or –OH groups are defined as being hydrogen bond donors. Calculated log $P$ values may be determined using either a fragmental or molecular additivity approach depending on the nature of the dataset [Lipinski et al., 1997]. The Rule of Five does not definitively categorise all well and poorly absorbed compounds, although it is simple, fast, and provides a reasonable degree of classification.

## 1.3 Modeling Techniques in Pharmacokinetics

Modeling provides a means to describe and understand data. It can also be useful for predictive purposes. Using a mathematical model such as a set of equations, large volumes of data may be summarised to provide a simpler representation. Depending on the modeling technique, pharmacokinetic data may be analysed using the model constructed. Pharmacokinetic models are relatively simple mathematical tools that represent complex physiologic processes. Thus, insight into mechanisms involved in pharmacokinetics, such as distribution and elimination, may be gained. Models which have been adequately validated can then be used for predictive purposes [Bourne, 1995]. This may be useful, amongst other things, in aiding lead compound selection or for failing unsuitable compounds early during the developmental processes.

Different modeling approaches can be more or less useful for a given modeling task. It is essential, therefore, to select the most appropriate modeling technique for each situation. ANNs were explored in the present wok because of their demonstrated ability to develop predictive data-driven models. Other methods may be more useful if the aim is to, for example, focus on mechanistic relationships in pharmacokinetics. Nevertheless, the flexible nature of soft computing makes ANNs a potentially useful tool in predictive pharmacokinetic modeling. Such flexibility dictates that replicate experiments be performed to provide a measure of experimental precision.

A crucial aspect of modeling for predictive purposes is validation of the final model. Validation involves testing the ability of such a model to make predictions. An unvalidated model is only useful for the data it was constructed on. Thus, model testing using a cross-validation technique (Section 2.6.2) or independent testing compounds is often employed. Validation of a model provides a measure of its predictive ability and/or potential utility.

### 1.3.1 Non Structure-Based Methods

The focus of this monograph is on structure-based methods of modeling, however non structure-based methods will be mentioned briefly here for completeness. Three approaches that have been suggested for pharmacokinetic modeling include the compartmental approach, physiologically based methods, and model-independent techniques.

The compartmental approach is an empirical approach which is based on a simple compartmental model. These compartments have no strict physiological or anatomical basis. The body is represented by a number of theoretical compartments that communicate reversibly with each other. The compartment can represent a body volume or, just as easily, it could represent a chemical state such as the metabolite of a drug. This approach usually uses either one or two compartments. Compartments are loosely considered a tissue or group of tissues with similar blood flow and drug affinity [Cutler, 1978]. Since there is more mathematical than physiological relevance for the parameters obtained in compartmental pharmacokinetic models, they cannot be used to extrapolate between species or provide mechanistic information about drug pharmacokinetics.

Despite its simplistic nature, many useful quantities can be derived using this approach and by comparing predicted values with actual data. They are also useful when only plasma or blood concentration-time data are available without necessarily requiring tissue concentration data.

A physiologically based pharmacokinetic (PBPK) model identifies the compartments with actual body spaces. Such models are a great deal more complex than simple compartmental models. PBPK modeling incorporates physicochemical data as well as anatomical and physiological data from animals or humans to develop models for pharmacokinetic prediction [Grass & Sinko, 2002]. These models describe the mechanistic inter-relationships between ADME processes. Hence, they are more adaptable to clinical therapy and for changing situations. PBPK models can also be used for predictive purposes. PBPK models may provide useful information describing drug disposition and metabolism [Poulin & Theil, 2002] but require a large number of experimental parameters for model construction [Balant & Gex-Fabry, 2000]. They are, therefore, rarely used in early drug discovery or development.

Both compartmental and PBPK models require multiple data points from a single subject. In contrast, population pharmacokinetic modeling can use pooled data from multiple subjects. This is particularly useful when pharmacokinetic data is "sparse," that is, when only limited data is available. In addition to sparse data, rich data can also be used for population pharmacokinetic modeling either separately, or in combination with sparse data [Tett et al., 1998]. One of the main advantages of population pharmacokinetic modeling is that data is gathered from a number of sources so fewer samples per subject are required. Another advantage is that pooled data allows conclusions regarding inter-subject variability in pharmacokinetics to be drawn. Disadvantages are that separate models must be developed for each drug and models are only representative of the species in question. Therefore, population pharmacokinetic modeling is not often used in the early stages of drug development.

The model-independent approach is purely mathematical [Kuhle et al., 2005]. It avoids recourse to kinetic parameters that may not be valid, and models developed tend to be less complex. This approach is good for modeling ADME values but gives no physiologically relevant information about drug properties.

## 1.3.2 Structure-Based Methods

The fundamental assumption in pharmacokinetic modeling based on structural considerations is that changes in molecular composition and atomic arrangement are quantitatively responsible for changes in drug pharmacokinetics. Such an assumption is based on the success of early QSAR models which demonstrated the relationship between pharmacologic activity and molecular structure [Seydel & Schaper, 1981]. Most of the reported activity values were from *in vitro* experiments using isolated organs or enzyme preparations. The challenge in moving from cellular systems to whole-body systems was, and still is, considerable. It has been proposed that structural variations may have a more obvious effect on pharmacokinetic parameters than pharmacological activity since they may be controlled to some extent by the physicochemical properties of a molecule. For example, the crossing of a biological membrane may be related to the lipophilicity of a molecule and, since the structure of biological membranes is consistent throughout the body, a consistent relationship between lipophilicity and membrane permeability should be expected [Seydel & Schaper, 1981]. The complexity of living organisms dictates that no such simple correlation is apparent: although relationships can be approximated they may not entirely be explained by structure.

### *1.3.2.1 Multilinear Regression*

The first QSAR models related physicochemical characteristics of a molecule with activity using mathematical regression equations. The Hansch analysis, or linear free-energy related (LFER) approach works on the principle that physicochemical properties of a compound are additive and may be combined linearly to approximate activity. This can be summarised in the following manner (Equation 1-1):

$$y = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon \qquad \textbf{Equation 1-1}$$

where $y$ is the independent variable and may be a biological activity or property, $k$ is the number of independent variables, $\beta_1 \ldots \beta_k$ are the regression coefficients, and $\varepsilon$ is a constant. The coefficients are determined by means of multiple linear regression using the least squares method. The assumption in LFER modeling is that the different magnitudes of a biological activity or property within a compound series correspond to changes in the free energy of the compounds which occur when reactions or interactions take place. The gradations of both activity/property and free energy are supposed to be linearly related [Hansch & Fujita, 1964]. Free energy is difficult to determine in biological systems so constants representative of free energy can be used. These constants include rate constants, and steric, electronic, and lipophilic parameters.

One of the advantages of modeling using such simple equations (Equation 1-1) is that direct relationships between variables and the target activity/property values are evident. A positive coefficient corresponds to an increase in the target value whilst the opposite is true for negative coefficients. The absolute size of the coefficient indicates the magnitude or importance of the contribution of a particular variable. Thus, if all data are scaled appropriately, a large absolute coefficient indicates an important contribution by a particular descriptor whereas the smaller the absolute value of a coefficient the less of an influence that descriptor has on the target value.

A disadvantage with multilinear regression (MLR) is that in general about four or five compounds (patterns) at least are required for each variable used. This is a problem with small datasets since the number of descriptor variables available are thus restricted. Such datasets may not provide sufficient information to enable construction of a meaningful model. Furthermore, the relationship between the number of patterns and number of variables needs to be monitored in multilinear regression to avoid chance effects [Topliss & Edwards, 1979]. Another disadvantage is that drug data often contains correlated or skewed information that can lead to the construction of poor regression models [Butina et al., 2002].

**Table 1-1.** Human QSPkRs using multilinear regression.

| Pharmacokinetic Parameter | Drug Data | Equation Variables | Reference |
|---|---|---|---|
| Half life | sulfonamides | *log P* | [Seydel et al., 1973] |
| Epidermal absorption | aliphatic alcohols | *log P* | [Lien, 1975] |
| Protein binding | penicillins | *log P* | [Craig & Welling, 1977] |
| Volume of distribution | penicillins | *P* | [Watanabe & Kozaki, 1978a; Watanabe & Kozaki, 1978b] |
| Metabolism by monoamineoxidase | aliphatic amines and alcohols | log *P*, p*Ka* | [Kubunyi, 1979] |
| Clearances, mean residence time, volumes of distribution | β-adrenoceptor antagonists | p*Ka*, $K'_{SF}$ (octanol/buffer partition coefficient) | [Hinderling et al., 1984b; Hinderling et al., 1984a] |
| Clearances, mean residence times, volumes of distribution | non-congeneric compounds (17) | MW, intrinsic and alcohol solubilities, protein binding, distribution coefficient | [Herman & Veng-Pedersen, 1994] |
| Volume of distribution | Non-congeneric compounds (129) | Fraction ionised, electrotopological indices, electrostatic potential, log *P*, lipole | [Ghafourian et al., 2006][a] |

[a]MLR combined with GA

Early QSPkRs relied mostly on experimental variables to develop multilinear regression equations for a range of pharmacokinetic parameters. The most common variable used, log *P*, has been correlated with many different ADME parameters in both animals [Winningham & Stamey, 1970; Martin & Hansch, 1971; Seydel et al., 1980; Blakey et al., 1997] and humans (Table 1-1). Early QSPkRs were generally constructed using only small congeneric datasets. Experimental values were obtained for each study individually to ensure consistency of results. Animal data was much easier to obtain than human data so animal QSPkRs are more prevalent in the literature. Success of early QSPkR studies for prediction and drug development was limited due to the small number of drugs and types of descriptors used. In addition, models were developed to relate descriptor variables with pharmacokinetic parameters of drugs only present in the training set. Early models were generally not further validated which limited their usefulness.

One recent study, however, combined MLR and GA techniques for prediction of volume of distribution for a structurally diverse series of drugs [Ghafourian et al., 2006]. The final regression equations incorporated both theoretical and experimentally-derived descriptors and were tested using a leave-25%-out technique. Consensus model accuracy was determined to be 72.2% of predictions to have less than a two-fold error.

While obviously useful, log *P* does not provide all the information about a molecule necessary for construction of unlimited sorts of structure-property relationships. Additional physicochemical descriptors have been incorporated in models over time and eventually theoretical descriptors were included as well [Genty et al., 2001]. Even so, the limitations of multilinear regression were still apparent: numbers of descriptors needed to be controlled and validation of models remained a challenge.

#### *1.3.2.2 Artificial Neural Networks*

In comparison with multilinear regression, ANNs are more flexible, robust, and better at prediction [Butina et al., 2002]. ANNs are inherently nonlinear and adaptive systems that have demonstrated robustness with the often complex and noisy experimental data in the pharmaceutical field [Turner et al., 2003a]. Their introduction into the area of pharmacokinetics has been relatively late compared with other scientific and industrial fields. ANNs have found use in clinical monitoring to match pharmacokinetic profiles with pharmacodynamic effects [Minor & Namini, 1996], and for predicting patient creatinine clearance based on physiological variables [Herman et al., 1999]. In both these studies, ANNs produced results superior to regression or other modeling methods alone, and conclusions were that ANNs would be of particular benefit in those clinical situations. Similarly, another study used physiological and demographic data to predict the pharmacokinetics and pharmacodynamics of repaglinide, an oral hypoglycaemic agent [Haidar et al., 2002]. In addition to obtaining acceptable predictions, the ANN technique also allowed identification of significant covariates. It is important to note that ANNs do not present mechanistic information so other methods must be used should the mechanisms underlying pharmacokinetics be required. Even so, similar predictive models have been obtained using ANNs when compared with mechanistic modeling techniques [Nestorov et al., 1999].

ANNs have also been compared with conventional programs for construction of predictive population pharmacokinetic models. In comparison with the population pharmacokinetic modeling program NONMEM [GloboMax LLC, 1998], ANNs demonstrated lower absolute and prediction errors [Chow et al., 1997]. Modeling of the pharmacokinetic data was accomplished well by both methods, however, results presented were not validated against external data. A more recent study trained both ANNs and NONMEM on two thirds of a 622 point dataset and found that prediction of the remaining one third of the data was consistently superior using the ANN model [Tolle et al., 2000]. In both these studies, the same covariates were used for ANN and NONMEM model construction. Hence, there appears to be good potential for ANN application in population pharmacokinetic modeling.

QSAR studies have benefited from the use of ANNs for over a decade [Aoyama et al., 1990]. Following the progress of QSAR, development of QSPkRs using ANNs has also advanced. Both physicochemical and theoretical descriptors have been used successfully in ANN QSPkR studies, either individually or in combination [Ritschel et al., 1995; Agatonovic-Kustrin et al., 2008]. The trend has been towards completely *in silico* models as described earlier, since the speed associated with ANN models coupled with cheaper and more powerful computational methods has made this direction more feasible. A more detailed discussion of the application of ANNs in QSPkR modeling is given in Section 1.6.

## 1.4 Artificial Intelligence Systems

Soft computing methods have been used to varying extents in pharmaceutical research. ANNs have been the most popular due their intrinsic nonlinearity and characteristic robustness. They have been termed "universal approximators" since by varying network architecture it is possible to model almost any given situation [Haykin, 1994]. In addition to simply modeling a system, their use in prediction has also received growing attention. Multilayer perceptron (MLP) and radial-basis function (RBF) ANNs have been used more extensively than other ANN paradigms, and they will be discussed in the proceeding sections. The Kohonen topology-preserving map, otherwise known as self-organising map (SOM) is a form of ANN which may be used as a clustering tool. In theory, relationships in a multidimensional space are mapped onto the surface of a torus so that the Euclidean distance separating each point is equal. Sectioning one part of the torus and unbending it produces a cylinder. Sectioning the cylinder along its length allows it to be unrolled to form a rectangle. Hence, the original points now reside on a 2-dimensional (2D) plane. One advantage of self-organising maps is that target values are not required for initial model construction. Genetic algorithms (GAs) are evolutionary systems based on chromosomal recombination and selection. They can be used as stand-alone models however they are often used in combination with other soft computing methods such as feed-forward back-propagation ANNs to select optimal subsets of descriptors in QSAR/QSPR studies [Terfloth & Gasteiger, 2001].

### 1.4.1 Multilayer Perceptron ANNs

ANNs are mathematical models inspired by the structure of the biological brain. They are composed of many individual processing units or artificial neurons which are extensively inter-connected to form a network. Emulation of brain function was based on the hypothesis that the information in a brain resides in the strength of connections between neurons and not in the internal state of the neurons themselves [Bucinski et al., 2000]. Learning is simply the adjustment of the strengths associated with each connection. For this reason, connections between neurons are known as "weights." The neurons are connected to one another in parallel which provides their characteristic speed, robustness, and generalisation ability (Maddalena, 1998).

Multilayer perceptrons are of the feed-forward back-propagation class. They are composed of neurons organised into an input layer, one or more hidden layers, and an output layer (Figure 1-2). The number of input neurons is equal to the number of variables, and the number of output neurons is equal to the number of targets being predicted which in most cases is equal to one. The number of neurons in the hidden layer can be either one, for studies which parallel multilinear regression, or greater than one for studies dealing with nonlinear data [Aoyama & Ichikawa, 1991]. In each layer there is usually also an extra bias neuron which is not connected to the neurons in the previous layer. The bias neuron provides a magnitude adjustment to the input values so that they are in the correct range for processing by other neurons [Swingler, 1996]. Since the bias neuron is connected to each and every neuron in the hidden and output layers, the relationship between input variables and target output is not easily traceable. As mentioned previously, in multilinear regression the influence of each variable is proportional to the size of its coefficient whereas this is not necessarily the case with multilayer perceptrons. In order for the multilayer perceptron to make predictions it must first be trained as described in the General Methodology (Section 2.6.1.1).

Multilayer perceptron ANNs have been widely utilised in the pharmaceutical sciences. In a clinical setting, ANNs were used to monitor the pharmacodynamics of short-acting neuromuscular blockers [Lendl et al., 1999]. ANNs were chosen since they offered a fast and controllable mechanism for prediction without the need for more costly biopharmaceutical data. Compared with conventional closed-loop controllers, results using the ANN were encouraging, suggesting a potential use for this technique in this and other clinical settings. In another clinical study both pharmacokinetic and pharmacodynamic relationships were analysed using ANNs [Minor & Namini, 1996]. It was also suggested that ANNs may be useful for time-dependent modeling and aiding in the development and analysis of clinical trials.

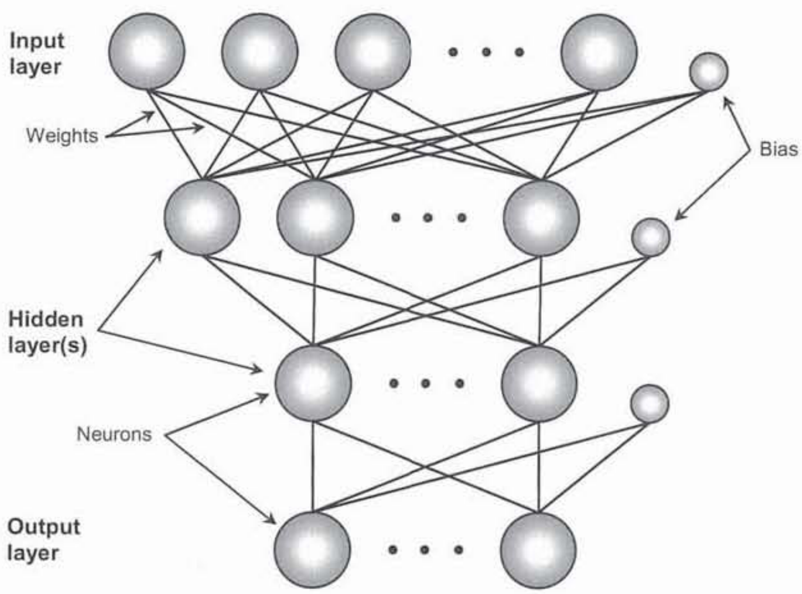



**Figure 1-2.** Schematic representation of a multilayer perceptron ANN.

Immunosuppressant therapy in organ transplant recipients requires close monitoring of drug concentrations to ensure adequate immunosuppression. Conventional methods rely on measurements of trough blood concentrations, although free plasma concentrations and two hour post-dose concentrations are also being examined. Using population data, clinical monitoring of peak and trough serum concentrations of gentamicin was examined [Brier & Aronoff, 1996]. Prediction of peak concentrations using ANNs was comparable with models constructed using NONMEM, while prediction of trough concentrations was superior using ANNs.

In other predictive applications the scaling-up of allometric data to predict human pharmacokinetic parameters has been performed with ANNs. In one study, animal data taken from the literature was used to predict human volume of distribution and clearance values [Hussain et al., 1993]. Although ANNs were shown to provide acceptable models, problems included the requirement of a substantial amount of training data which may not always be readily accessible. However, it was also shown that existing animal data was able to be supplemented with theoretical data from drug structure. Furthermore, drug physicochemical data may also be included for construction of pharmacokinetic models [Ritschel et al., 1995]. Prediction of clearance and volume of distribution using ANNs with such information was shown to be similar to *in vitro* estimations, so no superiority of either technique was apparent in that respect. However, time and cost savings gained using the ANN indicated that it was potentially a more useful technique. As with the majority of earlier studies using theoretical descriptors, little attempt was made to explain the relationship of such descriptors with the pharmacokinetic parameters in question.

### 1.4.2 Radial-Basis Function ANNs

Radial-basis function ANNs differ from multilayer perceptrons in that the nonlinear transformation of data occurs only in the hidden layer and not elsewhere [Yao et al., 2002b]. Radial-basis function ANNs belong to the class of kernel estimation methods and employ a transfer function representing a bell-shaped Gaussian response surface. In contrast, the transfer function in multilayer perceptrons is generally sigmoidal (Figure 1-3). Although they only have three layers of neurons available, radial-basis function ANNs are functionally comparable to multilayer perceptrons. Details of training are given in General Methodology (Section 2.6.1.2).

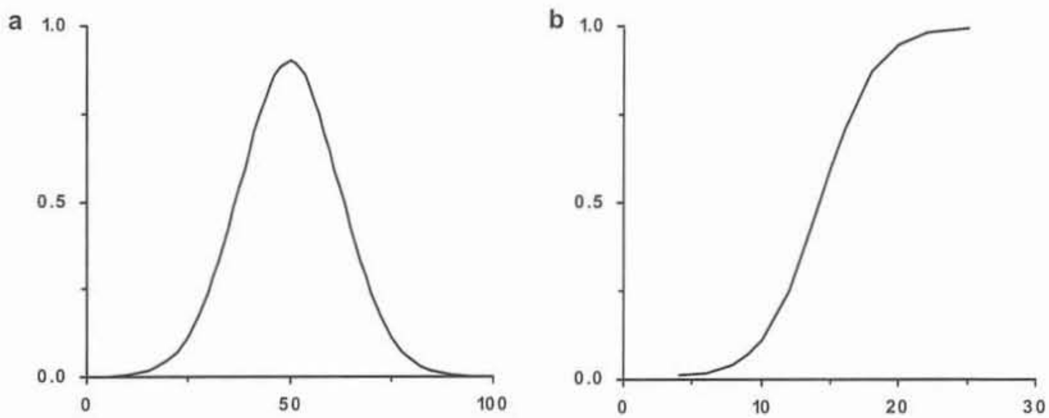



**Figure 1-3.** Transfer function for a) hidden neurons in a radial-basis function ANN, and b) neurons in a multilayer perceptron.

Radial-basis function ANNs have not been used as extensively as multilayer perceptrons in pharmacokinetics or elsewhere. For the same modeling task they generally require more neurons in the hidden layer than the latter which leads to increased network complexity. It has been suggested that complexity of a network can influence the training and predictive performance of a model. The analogy applied is that multilinear regression studies require a certain minimum number of patterns per optimisable parameter. That is true for any modeling technique, however, the analogy is not entirely applicable to ANNs in general since they are nonlinear systems. Hence, the relationship between the complexity of the model and the number of patterns depends specifically on the nature of the model itself [Turner et al., 2003a].

One advantage of radial-basis function ANNs over multilayer perceptron ANNs is the speed at which they are trained. In one QSPR study using 233 compounds both paradigms were directly compared [Tetteh et al., 1996]. It was found that results using both paradigms were comparable for both training and validation. However, radial-basis function ANNs trained faster and were less likely to fall into local minima than multilayer perceptrons that employed sigmoidal transfer functions. Further comparisons were also made with linear modeling techniques which were found not to be as useful as ANN models. A similar conclusion was drawn in other comparative studies, for example, in a recent QSPR for benzene derivatives employing quantum mechanical values as input descriptors [Wang et al., 2002].

Since radial-basis function ANNs train relatively fast, they are well suited to problems involving large datasets with numerous descriptor variables. Several studies have generated a number of theoretical descriptors and then selected only a subset to use in the final model. One such study generated 35 topological descriptors and selected a subset of 9 for the final QSPR of 173 compounds [Yao et al., 2002a]. Another QSPkR for intestinal permeability constructed at a predictive model of 15 descriptor variables from a total of 57 generated for 86 drugs [Agatonovic-Kustrin et al., 2001]. Both studies employed different methods for selection of optimum descriptors, but all models were suitably cross-validated to ensure soundness of results. The fact that radial-basis function ANNs require a greater number of hidden neurons than multilayer perceptrons does not usually affect training time significantly.

### 1.4.3 Generalised Regression Neural Networks

More recently there has been increasing interest in the use of generalised regression neural networks (GRNN) in pharmocokinetic modeling [Yap et al., 2006]. These also have the advantage of fast training but the disadvantage of often ambiguous descriptor interpretation. One direct comparison of GRNN, multilayer perceptron, and multilinear regression modeling of blood-brain barrier penetration of 159 compounds [Yap & Chen, 2005] utilised principle component analysis to select training and validation compounds, a genetic algoritm (GA) approach for descriptor selection, and further principle component analysis for descriptor analysis (see Section 1.4.5). This technique was extended to human seum albumin binding and milk/plasma partitioning models derived from 93 and 122 structurally diverse compounds respectively. Only theorectically-derived descriptors were included which, once explained by principle components, made interpretation of individual descriptors challenging. For all three datasets the GRNN approach yielded superior results although predictions and comparisons were not greatly convincing.

Another study [Niwa, 2003] used a known set of 86 compounds and their human intestinal absorption (HIA) values [Wessel et al., 1998] to test GRNN and probabalistic neural network modeling capabilities. Predictions were comparable to, although not as accurate as, the original HIA study. The methods employed proved useful in the speed, simplicity of model and descriptor generation, and interpretation of the final model. It remains, however, that improvements are necessary in terms of drug dataset size and quality.

### 1.4.4 Other Soft Computing Methods

Genetic algorithms are an evolutionary technique well suited for selection purposes. As such, they have been used to reduce the number of available compounds [Gobbi & Poppinger, 1998] and for compound selection optimisation to identify bioactive molecules [Gillet et al., 1998] in virtual libraries. As well as being a useful tool in isolation, genetic algorithms have often been applied in combination with other modeling techniques such as ANNs. These genetic neural networks (GNNs) provide an automated pruning technique for studies involving large numbers of descriptor variables [Zupan & Novic, 1999]. Comparatively, GNNs have matched QSAR results obtained using manual pruning, and have the added advantages of speed and unbiased descriptor selection [So & Karplus, 1996]. With the increasing number of theoretical descriptors able to be generated from drug structure, GNNs have also successfully aided the selection of key descriptors for QSPR models constructed using multilinear regression [Turner et al., 1998], and ANN [Agatonovic-Kustrin et al., 2001] methods. Although genetic algorithms present a useful alternative to manual selection of descriptors, they tend not to be used for exhaustive searching or correlating since they are computationally expensive relative to other *in silico* approaches. Nevertheless, GAs have recently been utilised in QSPkR studies as stand-alone models using simple theoretical descriptors [Wang 2006, Cheng 2007] with encouraging results.

A Kohonen self-organising map represents an unsupervised neural network paradigm, and is essentially a 2D representation of a multi-dimensional space [Kohonen, 1997]. With respect to pharmaceutical applications, a drug may be described by numerous physicochemical or theoretical descriptors and then represented by a position on a 2D map relative to other compounds. Thus, similar compounds are clustered together on the self-organising map whilst different compounds are positioned away from each other. This technique, known as k-nearest-neighbour (kNN), has been used to classify pharmacologically active molecules amongst non-congeneric datasets [Bauknecht et al., 1996] and also to locate potentially useful anticancer drugs [van Osdol et al., 2000]. Clustering of compounds according to odour properties has also been performed [Audouze et al., 2000] but it was found that appropriate description of the compounds for projection onto a self-organising map was problematical. Useful self-organising maps have been created in other areas such as protein surface [Stahl et al., 2000] and molecular surface potential mapping [Gasteiger et al., 1994]. One combined k-NN/simulated annealing study achieved predictive results for *Vss* and CL comparable to other soft computing methods, and superior to PLS modeling [Ng 2004].

Fuzzy logic is characterised by a mathematical framework that lacks well-defined boundaries. The lack of strict boundary conditions allows flexibility in classification problems, and fuzzy sets have successfully been developed to classify compounds according to their chemical composition [Pop et al., 1996], and similarity [Maggiora & Mezey, 1999]. In pharmaceutics, fuzzy sets have been applied in clinical drug dosage monitoring [Kern et al., 1997; Shieh et al., 2002], and prediction of serum pharmacokinetics [Sproule et al., 1997]. When compared with conventional population pharmacokinetic modeling using the same data, fuzzy logic provided comparable results and allowed determination of important covariates. Other pharmacokinetic studies have used fuzzy methods to predict human bioavailability and volume of distribution [Hirono et al., 1994a; Hirono et al., 1994b]. These studies grouped compounds according to known bioavailability values and further subdivided groups according to broad chemical composition. Even though fuzzy logic appears to be a useful alternative to conventional modeling, further development is required to enable prediction of unknown compounds.

### 1.4.5 Descriptor Selection

Optimum descriptor selection remains a fundamental problem in QSAR/QSPR studies. Many descriptors may be considered for inclusion in a model, however not all may provide useful information and indeed some may even be detrimental. The aim of descriptor selection is to improve model generalisation by reducing unnecessary data [Tetko et al., 1998].

Early multilinear regression models utilised smaller numbers of physicochemical descriptor variables such as log $P$ and association constants. As the field of QSAR/QSPR grew, additional ways of describing compounds were realised and methods had to be established to select relevant descriptors. Since multilinear regression provides a direct relationship between a descriptor and the output space, normalised descriptors with very low coefficients are considered not to contribute significantly to a given model and so may be removed without much harm. More complicated stepwise regression techniques have also been implemented. These involve identification of an initial model and then repeated alteration of the model from the previous step by the addition (forward stepwise) or removal (backward stepwise) of a descriptor variable. The search is terminated when stepping does not further improve the model.

Partial least squares regression is an extension of multilinear regression, but derives factors from the descriptor variables to maximise the covariance between the descriptor and output spaces [Bjork & Danielsson, 2002].

Clustering methods group similar descriptors together to minimise the variance within clusters but maximise variance between clusters. From these clusters, suitable descriptors may then be selected which should represent a substantial portion of the information contained in the entire descriptor set.

Principal component analysis (PCA) derives descriptors representative of the whole descriptor space but does so by linearly combining variables to maximise variance between the individual principal components [Bruni et al., 2002]. All these descriptor selection methods have been used to some extent in pharmaceutics [Abuzaruraloul et al., 1998] and QSAR , although other genetic algorithm and ANN techniques have proven more effective [Winkler et al., 1998].

As described earlier, genetic algorithms provide an effective means of descriptor selection which may then be combined with, for example, ANN modeling [So & Karplus, 1996]. Chromosomes, representing the entire descriptor space, are composed of genes, which represent individual descriptors, and randomly crossed-over to simulate biological evolution. A fitness function applied to the resultant offspring retains the better-performing chromosomes, and then the process is reiterated until the chromosome with the best genetic composition has evolved [Zupan & Novic, 1999]. Analogous to biological systems, mutations are sometimes incorporated to help offspring avoid local minima. One limitation with genetic algorithm searching is that chromosomes are often constrained to a fixed length, which may restrict the characteristics of the terminal offspring. Variable-length chromosomes have been used, however, this attenuates the problem of lengthy training times [Yasri & Hartsough, 2001].

In contrast, studies using ANNs alone are computationally inexpensive. Such studies often include the entire descriptor set initially followed by descriptor selection being performed on a continuous basis until an optimum subset is achieved. Various selection or pruning techniques exist which aim to eliminate redundant weights and/or descriptors, thus leaving only those offering a significant contribution to the model.

Pruning may be divided into sensitivity and penalty term methods, and can be implemented either manually or incorporated within the training algorithms. Sensitivity-based methods have been examined and it has been found that simpler magnitude-based algorithms performed as well as more sophisticated error-based algorithms [Tetko et al., 1996]. Penalty terms were also shown to be useful in removing redundant weights and were thus able to accentuate the importance of certain descriptors with respect to the target output space. Manual selective pruning has been used successfully to reduce the number of descriptors and, although time-consuming, has the advantage of allowing greater control over the pruning process than automatic algorithm-based techniques [Maddalena & Johnston, 1995].

## 1.5 Descriptors Used in Modeling

Classical physicochemical descriptors such as log *P* are not available for all known chemical entities. Conversely, theoretical descriptors may be calculated for all chemical entities should the structure be known [Devillers, 1999]. Descriptors may be a scalar representation such as atom-counts, or rely on a matrix, for example topological indices, or require lattice-type information to allow calculation of 3D descriptors. There are currently over a thousand theoretical descriptors that have been applied to chemical- and drug-related problems. Many theoretical descriptors contain similar information and this is particularly true for descriptors derived for smaller molecules or for structurally related compounds. It is generally accepted that appropriate methods should be undertaken to limit the number of topological indices in a study to those containing independent and useful information [Basak et al., 2000a]. Hence, suitable clustering or pruning is required to ensure that redundant descriptors are not included in the modeling process.

Independent descriptors are those which are not significantly linearly correlated with one another, thus, their information content is independent of that of other descriptors. To maintain diversity of information highly correlated descriptors are often excluded from a model. Even so, correlated descriptors may still be included in a successful models since, unless identical, they all contain a certain amount of independent and possibly useful information [Consonni et al., 2002].

Some of the more important theoretical descriptors and those relevant to the studies presented here will be described in the proceeding sections.

### 1.5.1 Constitutional Descriptors

Constitutional descriptors are the simplest of all theoretical descriptors, although they are not strictly classed as topological indices. They encode basic information such as the number and type of atoms in a molecule, and they also include counts of functional groups.

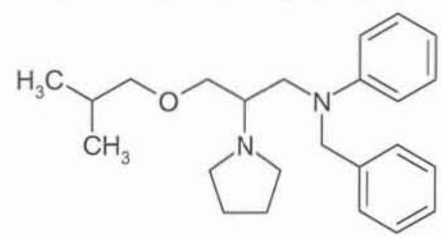
There are two ways of presenting constitutional descriptors: the first is using a binary system where the presence or absence of a particular moiety is denoted by a one or zero respectively. The second and more common method quantifies the number of cases of each moiety. For example, bepridil is represented in a different manner according to each system (Table 1-2). The assumption behind constitutional descriptors is that variations in atomic and functional group composition influence whole molecule properties. Indeed, this assumption underpins QSAR studies for drugs defined by a common template structure and altered at various substituent positions around that template [Maddalena & Johnston, 1995].

The fact that constitutional descriptors are easily interpreted is a distinct advantage over more complex topological indices. Constitutional descriptors have been used for prediction of physicochemical parameters [Burden, 1996] through to entire QSPkR analyses in combination with other descriptors [Agatonovic-Kustrin et al., 2001], wang 2006]. For prediction of biological activity, the ability of constitutional descriptors to encode useful information appears to be better suited to congeneric series of compounds since deviations in activity can be directly attributed to variations in substituents [Jaen-Oltra et al., 2000].

Application to non-congeneric series of compounds has been performed, although, to account for the larger differences in molecular structure, usually more complex descriptors are required as well. These additional descriptors may also be constitutionally-based such as the *V3* and *V4* indices which denote vertices of valence three and four respectively, and *L*, which is defined as the topological length of the two most separate points on the graph [Galvez et al., 1994b]. Other constitutional descriptors account for features such as number of rotatable bonds and molecular mass derivatives, and these have been included in models of drug solubility characteristics [Jorgensen & Duffy, 2002] and other pharmacokinetic parameters [Herman & Veng-Pedersen, 1994] with reasonable success.

**Table 1-2.** Constitutional encoding of bepridil ($C_{24}H_{34}N_2O$).

**Structure**



| **Encoding** | H | C | O | N | S | aromatic ring | carboxylic acid | quaternary ammonium | ether |
|---|---|---|---|---|---|---|---|---|---|
| **Binary** | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 1 |
| **Quantified** | 18 | 13 | 2 | 0 | 0 | 2 | 0 | 0 | 1 |

## 1.5.2 Topological Indices

Topological indices mathematically encode information regarding the structure of molecules which have been depicted as graphs. The molecular graph is comprised of vertices which correspond to atoms and edges corresponding to the bonds between these atoms. Often they are sensitive to size, shape, branching, cyclicity and, to a certain extent, electronic characteristics of molecules [Todeschini & Consonni, 2000]. Subgraphs are defined as two or more vertices connected by a bond or common path. Subgraphs may include branched and cyclic structures, and can have up to as many vertices as the entire molecular graph.

The seminal contribution to the field of topological indices was the introduction of the Wiener index. The Wiener index is defined as the sum over all bonds of the product of the number of vertices on each side of the bond [Wiener, 1947]. This index has been used extensively in the construction of QSPRs and QSARs for structurally related [Zakarya et al., 1993] and unrelated [Galvez et al., 1995] drugs. Performance has also been improved following modification of the single Weiner number to extended Weiner indices [Estrada, 1999].

The next significant topological index to be developed was the branching index proposed for a series of alkanes [Randic, 1975]. The Randic branching index, a precursor of the Kier and Hall connectivity indices, has led to successful elucidation of numerous topological-based QSARs and QSPRs. However, the physicochemical significance of the Randic index was undefined for decades after its inception. Only recently has the link between these theoretical numbers and their relation to physical chemistry been revealed. Research demonstrating that the branching corresponds to the relative area of accessibility of a molecule has established that physical meaning can be extracted from theoretical descriptors [Estrada, 2002a].

### *1.5.2.1 Connectivity Indices*

Kier and Hall connectivity indices, also called chi ($\chi$) indices, were developed to calculate zero- and higher-order connectivity descriptors [Kier & Hall, 1977]. Numerous correlations between connectivity indices and both physicochemical properties [Reinhard & Drefahl, 1999] and biological activity [Kier & Hall, 1986a] of drugs have been identified, mostly for structurally related compounds.

These descriptors are theoretical in nature so the absolute meaning of each index is not easily described. However, since the method of calculation is well defined then certain indices may describe some specific features of a molecule (Table 1-3). Moreover, recent work for both structurally related and structurally unrelated compounds has revealed the relationship between connectivity indices and molecular accessibility area [Estrada, 2002b]. Since molecular accessibility area is important in chemical interactions then such a relationship demonstrates the relevance of topological indices in drug models. Connectivity indices mathematically describe molecular structure by encoding branching and cyclicity (nonvalence $\chi$) and heteroatom influence (valence $\chi$). They cannot, however, encode absolutely every single structural feature of a molecule. For example, structures exhibiting *cis/trans* isomerism and atomic chirality are not differentiated from one another by connectivity indices [Cao & Yuan, 2002]. Modification of connectivity indices has been proposed to account for such shortcomings [Basak et al., 2000c], although simple connectivity indices still remain popular in the broader scientific and industrial community. Details of the methods of calculation of connectivity indices are given in Section 4.2.2.1.

**Table 1-3.** Information content of Kier and Hall connectivity indices [Kier, 1987].

| Index | Information Content |
|---|---|
| ${}^{0}\chi$ | General features about atoms or points, including molecular volume, molar refractivity, density and magnetic susceptibility |
| ${}^{1}\chi$ | Number of atoms in the molecule, and related surface area and volumes, relative branching in structural isomers |
| ${}^{1}\chi^{v}$ | Molar refractivity, orbital electronegativity, molecular polarity, structural differences for six-membered rings |
| ${}^{2}\chi$ | Information about branching (three-atom fragments) |
| ${}^{3}\chi_{p}$ | Flexibility, conformational *gauche-anti* rearrangements |
| ${}^{3}\chi_{c}$ | Branching, density, multiplicity of "cross-road" atoms |
| ${}^{4}\chi_{pc}$ | Structural description of substituted aromatic rings and information about the orientation of ring substituents |
| ${}^{4}\chi^{v}_{pc}$ | Number of benzene ring substituents, the substitution pattern, length of the substituents up to three bond lengths, and heteroatom type of substituent (in conjunction with ${}^{4}\chi_{pc}$) |

The application of simple connectivity indices extends from lead compound searching [Casaban-Ros et al., 1999] to QSARs and QSPkRs [Cercos del Pozo et al., 1996]. Studies have been performed for structurally diverse drugs and individual indexes have been shown to be important for different activity and property parameters. Linear combination of connectivity indices, for example differences and quotients, can describe features such as number and nature of heteroatoms, as well as inductive and mesomeric effects of molecules [Galvez et al., 1994b]. Hence, they provide a valuable adjunct to the information presented by connectivity alone and have been utilised successfully in structure-pharmacokinetic studies [Rose et al., 2002].

#### *1.5.2.2 Electrotopological Indices*

Biological and chemical properties of molecules rely on both their structural and electronic attributes. Since topological indices such as the Wiener index and connectivity indices describe features of a molecule principally from a structural perspective, topological charge indices were developed to explicitly describe the charge distribution characteristics in a molecule.

The topological charge indices, $G_k$, encode the total charge transfer between atoms in a molecule at a distance $k$ from one another [Galvez et al., 1994a]. Thus, $G_k$ indices are related to the dipole moment of a molecule, and can be of the order one to $L$. For acyclic compounds, $J_k$ indices represent the mean value for the charge transfer across the molecule and are a modification of the corresponding $G_k$ indices. Details regarding calculation are given in Section 4.2.2.1. Although originally defined for acyclic alkanes, $G_k$ and $J_k$ have been used to model physicochemical properties and biological activity of both structurally similar and structurally diverse drugs, including cyclic compounds [Galvez et al., 1994b; Galvez et al., 1995].

Electrotopological state indices, $S_i$, are based on the "intrinsic state" of atoms which is related to their valence state [Kier & Hall, 1990]. Intrinsic states have been defined for 39 different atom valence states, which allow calculation for a wide range of molecular structures. In addition to structure-activity relationships, these electrotopological state or E-state indices have successfully been correlated with aqueous solubility [Huuskonen et al., 1998], blood-brain partitioning in combination with connectivity indices [Rose et al., 2002], and volume of distribution [Ghafourian et al., 2006].

### 1.5.3 Quantum Chemical Numbers

Typical quantum chemical numbers include energies of the lowest unoccupied molecular orbital (LUMO) and highest occupied molecular orbital (HOMO), dipole moment, dielectric energy, steric energy, total energy, minimum energy, heat of formation, and electron affinity. Quantum chemical calculations rely on the 3D structure of molecules. Descriptors obtained in this manner are therefore sensitive to conformational changes of a compound. Molecules typically undergo an energy minimisation routine *in silico* in order to obtain the anticipated *in vivo* 3D conformation. Even though quantum chemical numbers provide information representing absolute thermodynamic and electronic properties for a molecule, these values may not apply to an *in vivo* situation where the conformation of the molecule differs from the proposed *in silico* representation. Thus, caution should be exercised when interpreting the meaning of quantum chemical numbers at a detailed level. They have, however, provided useful information on a broader scale from complexation [Estrada et al., 2001] to structure-pharmacokinetic studies [Ekins & Obach, 2000]. Moreover, it is easier to interpret the physicochemical meaning of quantum chemical numbers than it is to interpret other topological indices.

Several studies have demonstrated the importance of quantum chemical considerations in the intestinal permeability of structurally diverse compounds. One study examined 18 theoretical and quantum chemical descriptors using principal component analysis [Winiwarter et al., 1998]. It was established that the information content in those principal components was sufficient to indicate a relationship between structure and permeability. It was found, however, that quantum chemical numbers did not rank as highly as other theoretical partitioning and solubility descriptors, and that better models were constructed without the quantum chemical numbers.

Another study generated 42 theoretical descriptors for 254 drugs, and reduced the number to 12 in the final model. Of those 12, five were quantum chemical numbers, and their influence on membrane penetration was quantified [Agatonovic-Kustrin et al., 2001]. It was found that the significance of dielectric energy was more than double the next most important descriptor. Alone, quantum chemical numbers have been correlated with metabolism properties in rats [Cupid et al., 1999]. The urinary excretion of a series 22 of benzoic acid analogues and their metabolites was modeled using linear regression. Several models were developed and reasonable prediction correlations were achieved. Owing to the complexity of metabolic pathways, development of structure-metabolism relationships is considerably difficult. It has been demonstrated that relationships can be developed for structurally related compounds. Therefore, the next challenge would be to extend models to include large numbers of structurally diverse compounds.

### 1.5.4 Solubility and Partitioning

As described earlier, solubility and lipophilicity are vital elements in determining the entry of drugs into the body via the oral route. Solubility characteristics can limit absorption from the GI tract while oil/water partitioning can affect drug distribution and binding to proteins. Many studies have correlated physicochemical as well as theoretical descriptors with experimental solubility and oil/water partitioning, and this is dealt with in Section 1.6.1. Calculated solubility and solubility-related parameters are often related to the ionic and electronic characteristics of a molecule. One scheme represents solubility as a combination of dispersion, polarity and hydrogen bonding values to give a vector in 3D space which describes a "radius of interaction" of a molecule [Hansen, 1967]. In contrast, calculation of partition coefficients is performed according to an additivity method. One approach for generation of calculated log $P$ (clog $P$) sums the contribution to lipophilicity at an atomic level [Viswanadhan et al., 1989], while another employs contribution of functional groups to lipophilicity [Hansch, 1979]. There have been improvements suggested for both of these methods [Wildman & Crippen, 1999]. Nevertheless, clog $P$ values obtained using either method have been validated using large numbers of compounds, and have proven useful in a broad range of structure-activity/-property relationship applications [Lipinski et al., 2001].

Understandably, studies examining intestinal permeability have found clog $P$ to be an important descriptor. One study developed a number of models using different combinations of descriptors and found clog $P$ to have large regression coefficients in several models [Winiwarter et al., 1998]. Another ANN study using numerous theoretical descriptors determined clog $P$ to have the greatest effect on the model [Agatonovic-Kustrin et al., 2001]. The importance of clog $P$ as in indicator of bioavailability is also apparent from its inclusion in the Rule of Five [Lipinski et al., 1997] (Section 1.2.3) which is utilised widely in the drug development industry.

### 1.5.5 Other Descriptors

A multitude of other descriptors exist, some with easily identifiable meaning and others more abstract. Geometrical and bulk descriptors provide information regarding the 3D characteristics of a molecule. Calculated surface area, molar volume, and solvent accessible area all depend on the particular conformation adopted by a molecule *in silico*. Hence, a suitable approach to determine the 3D conformation of molecules must be employed to ensure validity of descriptors. These descriptors have not been used to a great extent in the literature for QSPkR analyses due to their dependence on molecular conformation and their lack of relevance to QSAR studies. QSPkRs rely on the ability of drug molecules to be absorbed and excreted which in turn depends on the size and shape of molecules. For example, size and shape characteristics can affect glomerular filtration, and they have been shown to affect the rate of membrane permeability [Ghafourian & Fooladi, 2001]. Accordingly, geometric and bulk parameters are potentially more useful for QSPkRs than QSARs.

Many topological descriptors rely in a specific representation of a molecule as a graph or matrix. By changing the manner of representation, different descriptors encode for diverse theoretical characteristics of a molecule. It would be prudent, therefore, to include numerous theoretical descriptors in the initial stages of model construction to adequately represent the multidimensional nature of a molecule and its potential interactions in physiolocial systems.

## 1.6 Structure-Pharmacokinetic Relationships

The ANN modeling technique utilised in the present research has been employed extensively in QSAR analyses over the last decade. There are many examples of activity, toxicity, and carcinogenicity studies in the literature, however, the reader is directed to the following references [Maddalena, 1996; Basak et al., 2000b; Buchwald & Bodor, 2002; Greene, 2002; Mager, 2006] to avoid unnecessary discussion in this work. Other structure-property applications in pharmaceutics include formulation optimisation [Takayama et al., 1999], partition coefficient prediction [Huuskonen et al., 2000b], and infrared spectra analysis, and chromatographic retention modeling [Agatonovic-Kustrin & Beresford, 2000].

Early QSPkR model development relied on regression equations to correlate physicochemical properties of drug molecules with pharmacokinetics [Haj-Yehia & Bialer, 1989]. It was suggested that QSPkRs should be confined to structurally similar compounds in order to avoid the risk of encountering discontinuities in pharmacokinetic properties. For example, log $P$ for a set of sulfonamides can determine plasma protein binding but may not effectively represent protein binding of penicillins. It was also proposed that models explaining only 60% of the variance for a particular pharmacokinetic parameter could be deemed adequate for the purpose of providing useful information about a particular congeneric series of drugs [Seydel & Schaper, 1981]. Considering the high stakes resting on successful drugs reaching the market as well as the cost of development of apparent failures, such a poor figure may now no longer be acceptable.

Recently, more functional QSPkRs have been developed to model the pharmacokinetics of structurally diverse sets of drug data [Herman & Veng-Pedersen, 1994]. Although still employing linear regression techniques, these QSPkR studies demonstrated that prediction of pharmacokinetic parameters was not necessarily limited to congeneric series of compounds. The use of physicochemical descriptors was the first logical step due to their general acceptance in other QSAR and QSPR studies. Theoretical descriptors have also been considered in combination with physicochemical descriptors to account for diffusion characteristics which may influence pharmacokinetics [Herman & Veng-Pedersen, 1994]. Finally, QSPkRs constructed solely from theoretical descriptors have shown promise in locating biologically active compounds amid structurally diverse drugs and also in modeling distribution half life [Galvez et al., 1996]. The regression techniques employed allowed simple models to be constructed. However, it has been the use of more robust soft computing methods that has increased over time instead.

### 1.6.1 Absorption

Absorption at the site of administration can influence drug bioavailability. Drugs administered intravenously do not undergo absorption processes, whereas other routes of administration generally require absorption to occur before a drug is available to the body. For orally-delivered formulations in particular (including tablets and capsules), disintegration, deaggregation, dissolution, absorption, and first-pass metabolism all contribute to the bioavailability of a drug. For a drug to be absorbed it must first go into solution in order to cross biological membranes (Figure 1-4). For an orally-delivered drug that has poor dissolution characteristics the time spent at absorption sites in the GI tract may be insufficient for complete absorption to occur. In such a case residence time in the GI tract can be increased by slowing intestinal motility, however, bioavailability may still be limited by the drug solubility.

$$\textbf{Solid dosage form} \xrightarrow[\pm\ \text{deaggregation}]{\pm\ \text{disintegration}} \textbf{Drug} \xrightarrow{\text{dissolution}} \textbf{Drug in solution} \xrightarrow[\text{membrane}]{\text{transfer across}} \textbf{Absorbed drug in blood}$$

**Figure 1-4.** Sequence of events in drug absorption from solid oral dosage forms.

One study examining three different datasets containing congeneric drug compounds aimed to predict aqueous solubility using topological indices [Huuskonen et al., 1997]. Solubility data was taken from the literature and cluster analysis was used to select an uncorrelated subset of five descriptors from the many descriptors that were generated. Models were subjected to the leave-one-out (LOO) cross-validation testing to overcome over-training and also as an indication of predictive ability. Even though it does not provide a true measure of the predictive ability of a model LOO cross-validation is useful when the size of datasets is limited.

The results obtained demonstrated a number of important points. First, successful models were able to be constructed using simple calculated descriptors rather than from experimental data. Second, ANN models were found to be more robust than regression models of similar data and, third, some models required more than a single class of topological indices to enable reasonable prediction. In further developments, other topological indices have been used to construct models for drugs not part of a congeneric series. Since 3D structure is important in dissolution, descriptors encoding geometrical properties of a molecule as well as charge distribution have been correlated with aqueous solubility using both regression and ANN techniques [Bodor et al., 1991]. Since 3D information was required, energy minimisation routines were applied to drug structures to arrive at appropriate conformations. True predictive ability was tested with an independent set of compounds and in most cases ANN models were found to be superior to regression models. The requirement to fully represent all substituents and structure permutations in the training set was apparent with the finding of one outlier in the independent test set which was not completely represented in the training set.

Other recent structure-solubility relationship studies have examined even more diverse drug datasets inclusive of compounds containing heterocyclic rings and multiple functional groups [Huuskonen et al., 1998]. Large numbers of topological indices were generated, and sensitivity-based pruning was applied to determine the most influential descriptors. Although the approach and results were sound, applicability to a broader range of chemical structures was restricted because of the limited structure representation in the original dataset. By increasing the size of the drug dataset, a greater variation in structure was represented. In addition, larger test sets were able to be employed to provide a better estimate of predictive ability of the model [Huuskonen et al., 2000a]. Predictive results were improved but since only 2D electrotopological indices were employed the model was not able to be explained in physical terms.

Once in solution, the penetration of drug molecules across a membrane is the next step in absorption. One study modeled passive drug absorption in rat intestine for a small, structurally diverse series of drugs using immobilised artificial membranes [Genty et al., 2001]. Consistent experimental methods and conditions were ensured by performing *in vitro* experiments to determine the input descriptors and target absorption values rather than collecting data from the literature. However, the resources required for such experiments were greater than needed for pure *in silico* modeling. At any rate, improved predictive ability was obtained with the addition of a theoretical descriptor.

Similarly, a combination of experimental and theoretical descriptors has also been used for prediction of human intestinal absorption of drugs [Winiwarter et al., 1998]. The limitation of experimentally determining log *P* as a descriptor variable meant that only a small training set of compounds was used. The variation in structure was assumed to be representative of a large number of current drugs so prediction of the absorption of independent compounds was reasonable. A larger study using ANNs developed cross-validated models which were then tested with independent compounds [Wessel et al., 1998]. Only theoretical descriptors were utilised in model construction, with absorption data taken from the literature. A number of methods were used to reduce the 162 descriptor set to the final 6, all of which eliminated rather than combined descriptors. Since many of the descriptors were linearly correlated, different final combinations could be found using genetic algorithm selection. Use of cross-validation and independent test sets reduced the training set to 76 compounds for which the model constructed could not be considered a broadly applicable predictive tool. The model did, however, clearly differentiate between drugs with high and low absorption values.

A similar study approached prediction of intestinal absorption with a view to explaining the meaning of theoretical descriptors in the QSPkR model [Agatonovic-Kustrin et al., 2001]. This was achieved using a radial-basis function ANN and by generation of descriptors encoding atomistic to 3D holistic properties. There were 15 descriptors in the optimum model representing a combination of constitutional, hydrophobic, electronic and steric properties. In addition to qualitatively indicating absorption characteristics, predictive results were quantitatively more accurate than in the original study. It was emphasised that studies based on literature data should be selective to avoid accumulation of poor or inappropriate data.

A number of structure-bioavailability relationships have been established using theoretical descriptors. One model constructed for 232 commercial drugs classified compounds into four classes according to their predicted bioavailability [Yoshida & Topliss, 2000]. Another QSPkR for 591 compounds developed using stepwise regression demonstrated that predictions were more accurate than those achieved using Rule of Five [Andrews et al., 2000]. Both studies included compounds spanning a broad range of chemical structures which made them substantially more valuable than models constructed simply from structurally related compounds. Prediction of bioavailability, as opposed to broad classification, was performed in the latter and not the former, whereas model testing using independent compounds was performed in the former and not the latter. To be of most use in drug development, models should ultimately contain aspects of both and be developed to quantitatively predict the bioavailability of unknown compounds.

### 1.6.2 Distribution

Once a drug is absorbed into the systemic circulation its subsequent reversible transfer to extravascular fluids and tissues is termed distribution (Figure 1-5).

Distribution is usually a more rapid process than elimination such that distribution is generally complete while there is still an appreciable amount of drug in the body. Drugs are often bound to proteins in the plasma such as albumin or in tissues. Some drugs distribute preferentially to tissues such as muscle, brain, skin, and fat, or to organs involved in elimination such as the kidney or liver.

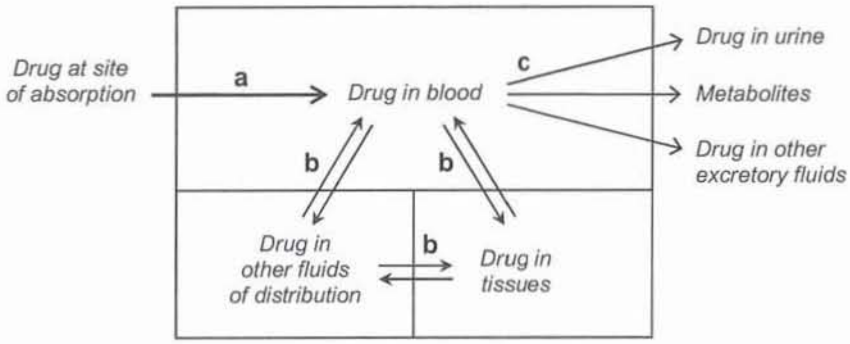



**Figure 1-5.** Schematic diagram of a) absorption, b) distribution, and c) elimination in humans (adapted from [Gibaldi, 1984b]).

The partitioning of a drug between tissue and blood is an important pharmacokinetic property describing the distribution of that drug in the body under steady-state conditions. Extent of partitioning is given by the partition coefficient, $P_{t:b}$, is defined as follows (Equation 1-2):

$$P_{t:b} = {}^{C_t}\!/_{C_b} \qquad \textbf{Equation 1-2}$$

where $C_t$ is the concentration of drug in the tissue of interest and $C_b$ is the drug concentration in the blood [Shargel & Yu, 1999]. Depending on the characteristics of the drug and availability of plasma pharmacokinetic data, it may sometimes be more appropriate to use the tissue/plasma partition coefficient, $P_{t:p}$.

Several studies have constructed models to predict $P_{t:p}$ from experimental oil/water partitioning and protein binding measurements [Poulin & Theil, 2000; Poulin et al., 2001]. Models were applied to structurally unrelated compounds for a range of tissues in rabbit, rat, mouse and human. The mechanistic nature of the models allowed deductions regarding the effect of lipophilicity on partitioning, as well as causal factors for distribution to particular tissues. Progress was made towards the goal of developing *in silico* prediction tools from literature data, however the drug dataset size limited the general applicability of these models to more diverse chemical entities.

A comparison of mechanistic and ANN methodologies demonstrated that both techniques were able to provide acceptable models for prediction of log *P* and tissue-to-unbound plasma concentrations for series of analogues [Nestorov et al., 1999]. Physicochemical data was determined experimentally for the construction of both models, however, suitable literature data could have been used instead. Both ANN and PBPK models were constructed using the same descriptive data. The ANN model provided similarly accurate predictions as the PBPK model did but did not supply any mechanistic information. Should predictions only and not mechanistic information be required, the ANN model could be considered to have equal performance to the PBPK model. Neither of the models was deemed superior and it was suggested that the alternative technique should not be discarded in favour of the other. Instead, they should be used to complement one another.

In another structure-distribution study the distribution of a broad range of drugs into the brain was modeled [Basak et al., 1996]. Input descriptors were topological and E-state indices and the target output parameter was the blood/brain partition coefficient. Variables were manually pruned initially and then subjected to statistical analysis to develop regression equations to predict the blood/brain partition coefficient. In addition to demonstrating their importance in determining tissue distribution, the optimum three-descriptor model allowed the relationship between topological indices and the physicochemical parameters of hydrogen bonding, aromaticity, and molecular branching to be examined.

The apparent volume of distribution does not have a true physiological meaning but represents the theoretical volume into which the drug is distributed. Volume of distribution at steady state can be defined as (Equation 1-3):

$$V_{ss} = {A_b}/{C_p}$$ **Equation 1-3**

where $A_b$ is the amount of drug in the body and $C_p$ is the concentration of drug in the plasma. High volumes of distribution indicate the preference of a drug to reside in tissues outside the plasma including erythrocytes and extravascular tissues. Conversely, low volumes of distribution indicate that a drug is confined mainly to the plasma [Rowland & Tozer, 1995d]. Apparent volume of distribution, or more specifically the volume of distribution of the unbound fraction, can provide useful clinical information since it is generally considered that the unbound fraction is responsible for the pharmacological action a of drug.

Prediction of volume of distribution has been performed using ANNs for 45 structurally unrelated drugs [Ritschel et al., 1995]. Physicochemical parameters of compounds, allometric data, and theoretical descriptors were used as model inputs. Validation was performed using a leave-*n*-out method. Rather than selecting an optimum set of descriptors, various combinations were evaluated for predictive ability. It was found that models which included log *P*, protein binding, and allometric data performed the best, although no detailed analysis of descriptor significance was performed.

A similar study constructed various QSPkR models for a small series of β-adrenoceptor antagonists [Gobburu & Shelver, 1995]. Volume of distribution of the total drug at steady state as well as the volume of distribution of the unbound fraction at steady state were modeled using ANNs and physicochemical descriptors taken from the literature. Only one neuron in the hidden layer of the ANN was required to model the volume of distribution of the unbound fraction. This indicated a linear relationship between the input and output spaces. Fraction bound to plasma proteins was also modeled and required a relatively large number of hidden neurons. Since ANN architecture is generally representative of the complexity of the parameter being modeled it was surprising to see the fraction bound requiring five hidden neurons when compared with the single hidden neuron model for volume of distribution of the unbound fraction. Nevertheless, ANN results were an improvement on those obtained using multilinear regression for both training, LOO cross-validation, and testing set predictions.

### 1.6.3 Metabolism and Excretion

Elimination of drugs for the body occurs via the processes of metabolism and excretion (Figure 1-6).

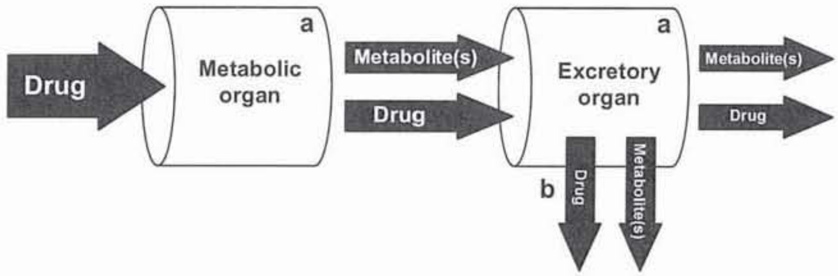



**Figure 1-6.** Drug elimination depicted schematically, showing processes of a) metabolism and b) excretion (adapted from [Rowland & Tozer, 1995b]).

Drug metabolism, or biotransformation, is the chemical conversion of the parent species into one of a number of metabolites. Drugs are excreted from the body as either the parent compound or as one of the metabolites. Excretion is primarily via the renal pathway via the urine, and also through the hepatobiliary route via the faeces. For volatile compounds it may be through the breath, and occasionally drugs are also excreted in the sweat. These are mainly Phase I reduction, oxidation, or hydrolysis reactions and/or Phase II conjugation reactions. Enzymes involved in metabolism are located predominantly in the liver, although other sites of enzymatic metabolism include the intestinal wall, kidney, lung, and skin. The major class involved in metabolism is the cytochrome P450 (CYP) enzyme superfamily.

CYP isozymes often display structural specificity in metabolic reactions. Approaching metabolism from a modeling perspective is greatly challenging due to the complexity of the metabolic system in its entirety. Different drug classes can be metabolised by different enzyme classes, and the same is often true even for structurally related compounds. Moreover, it is not uncommon for a drug to metabolised by simultaneous and competing enzymatic pathways. The extent of such metabolism depends on the individual rate constants of the competing pathways. To further complicate matters, expression and morphology of metabolic enzymes can show large variations across ethnic groups.

One approach in structure-metabolism relationship modeling has been to characterise molecules or structural motifs likely to interact with one specific enzyme [Mlinsek et al., 2001]. In doing so, the problem is simplified to dealing with a known target structure composed of one or more identifiable active sites. Thus, characterisation of the structural characteristics of theoretical ligands which affect binding is more easily performed. A study examining flavonoid derivatives made use of quantum chemical descriptors to model the inhibition of CYP 1A2 [Moon et al., 2000]. 3D conformation of ligands is crucial for enzyme binding so all structures were presented as energy-minimised conformers. Both multilinear regression and ANN models were constructed, with one ANN model making use of the same descriptors as the multilinear regression model. Direct comparison of predictive ability of each technique was thus possible, and since nonlinear relationships were assumed in flavonoid-CYP 1A2 binding, the ANN provided superior results. Another subset of descriptors was determined for the ANN using sensitivity-based pruning. This subset allowed more accurate predictions for the test set of compounds. This finding demonstrated the need for nonlinear methods of descriptor selection to be employed for nonlinear systems.

A more difficult approach is to model the enzymatic biotransformation of drugs by possibly numerous potential enzyme species. Complexity can be reduced by examination of just a single enzyme class instead of looking broadly at multiple enzyme classes. Carboxylic ester hydrolases catalyse the hydrolysis of a variety of ester-containing substances and are present in many human tissues. They have broad substrate specificity towards esters and amides and it is known that a single drug compound can be hydrolysed by more than one particular enzyme. The *in vitro* metabolism of a number of structurally unrelated compounds was modeled based on structural characteristics [Buchwald & Bodor, 1999]. Conformation of substrate drugs was determined by rigorous *in silico* minimisation routines. From the minimised structures quantum chemical numbers, calculated log $P$ descriptors, and a novel descriptor representing a theoretical steric angle were generated. Even though training results were acceptable the prediction of enzymatic metabolism of independent test compounds proved more difficult. The final models qualitatively differentiated between slowly and quickly metabolised substrates but quantitative predictions varied considerably from experimental values. An accurate predictive model was not expected, however, due to the size of the training dataset and complexity of enzymatic metabolism.

In a different problem, classification of metabolic fate of drugs has been performed for structurally related compounds. One study examined urinary excretion of glucuronide conjugates, glycine conjugates, and unchanged parent drug for 22 benzoic acid derivatives [Cupid et al., 1999]. Information regarding the metabolic enzymes was neither required nor explored since the study was not mechanistic by nature. Descriptors were calculated from structure and included quantum chemical numbers, geometrical descriptors, and partition coefficients. Predictive capability determined using a leave-2-out procedure indicated that structure-metabolism relationships were able to be modeled for the structurally-related compounds examined.

#### *1.6.3.1 Clearance*

Clearance is defined as the volume of blood cleared of drug per unit time. It is a function of both the intrinsic ability of eliminating organs such as the liver and kidney to excrete or metabolise a drug, and the blood flow rate to these organs (Figure 1-7). Clearance due to a single organ is given as the product of the blood flow to that organ and the extraction ratio (Equation 1-4):

$$CL = Q \cdot \frac{C_A - C_V}{C_A} = Q \cdot ER \qquad \textbf{Equation 1-4}$$

where $Q$ is the blood flow to the organ, $C_A$ is the concentration of drug in the arterial blood, $C_V$ is the concentration of drug in the venous blood, and $ER$ is the extraction ratio. The extraction ratio is the ratio of the rate of elimination of a drug to the input rate of the drug to an organ. Thus, the higher the extraction ratio the more drug is eliminated and the less passes through the eliminating organ intact.

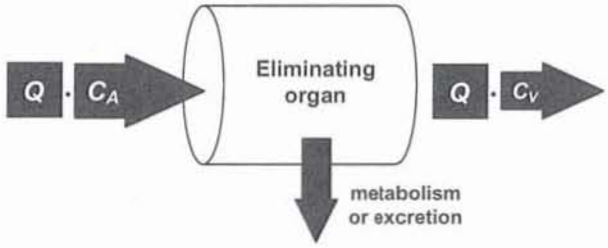



**Figure 1-7.** Schematic diagram of drug elimination by a single organ (adapted from [Gibaldi, 1984a]).

Total clearance is a complex parameter since it combines elements of both metabolism and excretion. A simpler parameter is renal clearance which mainly involves the processes of glomerular filtration and reabsorption. A comparison of models for renal and nonrenal clearance for a series of β-adrenoceptor antagonists gave similar cross-validation results for both [Gobburu & Shelver, 1995]. Even though this was not the expected result it was most likely due to the small size of the dataset. When clearances of independent compounds were examined, only the prediction of renal clearance was acceptable whereas prediction of nonrenal clearance was not.

In a related study a slightly larger dataset of more structurally diverse drugs was used to develop a model for *in vivo* hepatic clearance [Schneider et al., 1999]. The aim was to compare different modeling techniques to determine the feasibility of developing structure-clearance relationship models. A combination of allometric and *in vitro* data were used as input variables. It was found that adequate cross-validated models could be achieved using *in vitro* data only. Even though the dataset was relatively small a variety of metabolic pathways were represented covering both Phase I and Phase II reactions. Nevertheless, more broadly applicable models would require much larger datasets to cover the large range of metabolic pathways and also to allow true predictive performance using independent compounds to be assessed.

A structure-clearance relationship was developed in conjunction with the structure-distribution relationship mentioned in Section 1.6.2 [Ritschel et al., 1995]. The ANN model was novel in that two pharmacokinetic parameters, clearance and volume of distribution, were predicted simultaneously. Models were cross-validated but testing of independent drugs was not performed. Models were for the most part rather inaccurate. Moreover, quantitative structure-pharmacokinetic relationships were not examined. It was likely that the information content of the descriptors was inadequate to predict two complex pharmacokinetic parameters at once.

## 1.7 Summary Remarks

A review of the literature revealed that predictive QSPkR modeling is a relatively undeveloped area in which there is ample scope for progress using newer soft computing techniques such as ANNs. A number of methodologies have been applied from both mechanistic and non-mechanistic approaches. Should solely predictive models be required then a non-mechanistic approach would seem appropriate. To aid in drug development, however, information should be taken from the QSPkR models regarding the quantitative effects of structure on NCE pharmacokinetics. QSPkRs that have been developed to date have mainly used physicochemical or experimentally-derived parameters to construct models. Theoretical descriptors offer a quicker and more effective alternative. Each theoretical descriptor provides a certain amount of unique information. Modeling techniques need to be used that allow selection of the most appropriate set of descriptors for optimum predictive capabilities. Moreover, a combination of descriptors should be used initially to encode as much of the multidimensional nature of a chemical structure as possible.

ANNs have thus far only been used to a small extent in developing QSPkRs. They are a robust modeling tool and have certain speed and nonlinearity advantages where other methods do not. Construction of ANN-based QSPkRs should be demonstrated to be effective in simple systems first such as for structurally related sets of compounds, and then extended to include larger numbers of structurally diverse compounds. All models should be validated with an appropriate technique to ensure adequate predictive capability. In the literature to date most studies have addressed some of these issues but not all have to any great extent. The studies presented in this moograph validates the applicability of the present technique for both small and large sets of data, and in doing so addresses the issue of congeneric and structurally diverse drug datasets. The modeling described takes its basis from conventional QSPkR modeling published using both physicochemical and theoretical descriptors. All models were appropriately validated to ensure reliability in contrast to some earlier studies. In utilising ANNs and theoretical descriptors these studies have advanced the field of QSPkR modeling with a view to aiding the drug discovery and development process in the early stages.